Université Paris Diderot - Paris 7 2017-2018

Master 2 Biologie-Informatique/ Bioinformatique

Etude de la fonction et des mécanismes d'évolution des séquences répétées centromériques chez les Primates

Sarah Kaddah

Tuteur : Loïc Ponger

Structure et Instabilité des Génomes

MNHN - CNRS UMR 7196 / INSERM U1154 - Sorbonne Universités

Muséum national d'Histoire naturelle, 43 rue Cuvier 75005 PARIS

Remerciements

Merci à Namrod pour toute la partie sur la bibliographie. Retrouvez ses questions FAQ qui ont permis la rédaction de cette partie.
Merci à f-leb, LittleWhite et Metalman pour leurs conseils et la relecture. Merci à ced et jacques_jean pour la correction orthographique et typographique.

Table des matières

1 Introduction

1.1 Les séquences centromériques

-> biblio CENP-A

-> biblio kinetochore

-> info supp sur l'ADN satellite

Le centromère est une structure chromatinienne caractérisé par la présence de CENP-A. Cette protéine, très conservée au cours de l'évolution, est un variant de l'histone H3. Son rôle est de fixer la position du kinétochore par un mécanisme encore peu connu. En effet, le centromère est le site d'assemblage du kinétochore, un ensemble d'ADN et de protéines. Il permet l'attachement du fuseau mitotique pour la ségrégation des chromosomes durant la division cellulaire chez les eucaryotes. Le centromère et les protéines impliquées sont relativement bien conservés. Au contraire, l'ADN sous-jacent est très diversifié et l'organisation varie d'un taxon à l'autre. Cependant, une caractéristique commune est retrouvée chez toute les espèces : de l'ADN centromérique répété en tandem nommé ADN satellite. Ces répétitions sont issues d'événements d'amplification, tels les crossovers inégaux, la conversion de gènes, les cercles roulants ou la transposition de séquences.[Malik and Henikoff, 2002 ; Plohl et al. 2012] Ces séquences représentent 5% du génome. Les répétitions s'étendent de 7pb à 3,2kb avec des séquences de 145-180kb le plus souvent.

1.2 L'ADN α-satellites

-> première mise en évidence des AS

->théorie gradient de l'âge

-> travaux sur le gorille à dev

L'ADN satellite chez les Primates est connu sous le nom d'α-satellite. Ces séquences centromériques répétées en tandem sont riches en AT. -> article sur la première découverte

Des études chez l'Homme propose un modèle évolutif. La répartition des α-satellites suivrait une répartition spécifique selon l'âge des familles. Les familles les plus jeunes s'insèrent au cœur du centromère, repoussant les familles les plus anciennes jusqu'aux regions voisines, appelé péri-centromère.

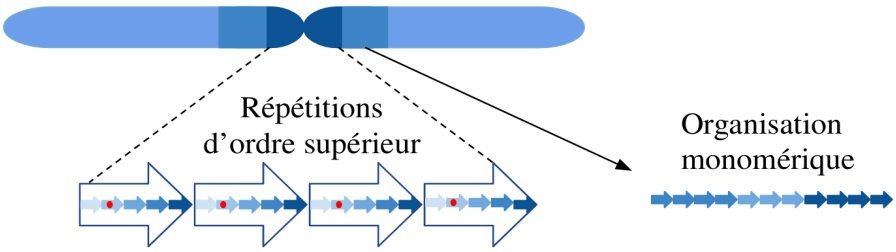

FIGURE 1 – **Organisation spatiale des α-satellites :** Le coeur du centromère (bleu foncé) est organisé en répétition d'ordre supérieur. Le péricentromère(bleu clair) a une organisation monomérique. Un monomère d'une même famille est représenté par une petite flèche de même couleur. Les points rouges représentent les sites de fixation à CENP-B ou pJα.

->Article Shepelev

->Est-ce que je peux utiliser du conditionnel ? OU est-ce que cette théorie est confirmée ?

Un monomère a une longueur de 171pb et il peut être répété des milliers de fois. Les monomères peuvent être répartis en famille selon leur similarité, les séquences ayant un taux d'identité supérieur à 70%. Ces séquences ont soit une organisation monomérique soit une organisation en répétition d'ordre supérieur (Fig. 1). Dans le premier cas, les séquences d'une même famille sont répétés en tandem. Dans le deuxème cas, une suite de monomères appartenant à différentes familles forme une unité, qui elle est répétée en tandem.

Ces séquences peuvent avoir un site de liaison à la protéine centromérique CENP-B un motif spécifique de 17pb. Cette protéine, qui reconnaît et se fixe sur l'ADN, serait présente chez de nombreuses familles de Primates. La protéine pJα, une protéine peu caractérisée, reconnaît un motif qui remplace celui de CENP-B.

Les α-satellites ont essentiellement été étudiées chez l'homme. Modèle évolutif avec les centromères en expansion. Une hypothèse concernant l'âge des séquences découle de ces recherches : les séquences les plus récentes apparaissent au coeur du centromères, déplançant les plus anciennes au péricentromère. D'autres études chez le gorille ont été faites. Le rôle des α-satellites est encore mal connu.

1.3 Le sujet de stage

enchaîne sur l'étude chez les cerco, une autre étude de séquençage haut débit

-travaux précédents limités (expliquer pk). Les méthodes basées aur l'alignement et la phylogénie sont très limitées, le jeu de données étant trop grand. Les méthodes n'étaient pas objec-

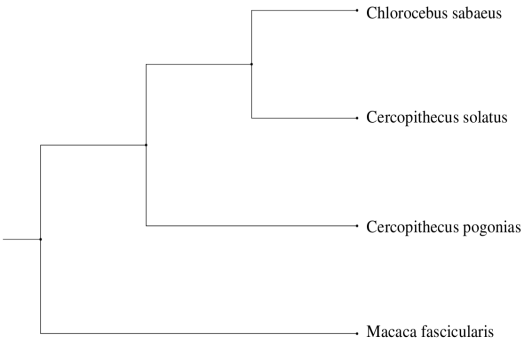

FIGURE 2 – **Arbre phylogénétique des espèces choisies.**

tives (quelles méthodes ??). De plus, chez d'autres espèces de Primates, les informations sont trop dispersées et aucune comparaison interespèce n'a été faite.

L'équipe d'accueil de mon stage "ADN répété, Chromatine, Evolution" ou ARChE, a récemment développé une approche de séquençage haut débit, ciblée sur les séquences α-satellites chez deux espèces de Cercopithèques. Une autre étude avec un grand nombre de séquences concerne le Gorille [Catacchio] avec l'utilisation de fragments relativement longs.

L'objectif de ce stage est de comprendre la fonction des α-satellites et leur mécanisme d'évolution. Je vais choisir plusieurs espèces de Primates. Je vais utiliser une méthode de classification automatisée améliorée du laboratoire [Florence Jornod] pour classer les séquences en familles. Ce programme permet de traiter des centaines de milliers de séquences sans quelque soit le nombre de séquences ou la taille des familles. Je vais dans un premier temps appliquer cette technologies au données issues de ce séquençage. Ensuite, je vais étudier d'autres espèces. Puisque toutes les espèces sont étudiées par la même méthode, une comparaison inter espèce est envisageable.

2 Matériel et méthode

2.1 Choix des espèces

Les critères de sélection dépendent de la disponibilité des séquences de qualité. Deux espèces du laboratoire sont choisies, les *Cercopithèques solatus* et *pogonias*, et deux espèces proches (Fig. 2), le *Macaca fascicularis* et le *Chlorocebus sabaeus*.

2.2 Méthode de classification

2.2.1 Principe

Cette méthode [1] répartit des séquences α-satellites en familles selon la similarité. La classification est hiérarchique dichotomique. Au départ, une table de 5-mers est calculée pour chaque monomère. Ensuite, toutes les séquences sont ajoutées dans la file. Ensutie une boucle itérative est exécutée pour séparer les séquences en groupes tant que les nouveaux groupes formés sont divisibles.

2.2.2 Répartition itérative

Une Analyse en Composante Principale (ACP) est effectuée sur la table de fréquence des 5-mers afin de réduire les dimensions du jeu de données et d'obtenir des variables indépendantes. Des distances euclidiennes sont calculées entre toutes les paires de séquence dans l'espace défini par les M premières composantes de l'ACP. A partir du calcul de distance, les séquences sont séparées en deux classes en utilisant la classification hiérarchique basée sur la méthode de Ward. La classification hiérarchique permet de former des classer de façon à maximiser l'inertie interclasse. Cette étape fait un usage important de la mémoire. Par conséquent, pour traiter des jeux de données importants, l'Analyse Discriminante Linéaire , une méthode d'apprentissage, est utilisée sur un sous jeu de données formé par des séquences tirées aléatoirement. Le modèle construit est alors appliqué sur toutes les séquences.

2.2.3 Validation d'un sous-groupe

L'étape suivante est une double validation des sous-groupes. D'une part la taille du sous-groupe est vérifiée. La taille minimale d'une famille est fixée à 100. Si un groupe atteint 100 séquences, il n'est pas redivisé. D'autre part les deux groupes doivent être distincts. Pour cela le *matepair*, la proportion de monomères ayant son plus proche voisin dans le même groupe, est évalué. Des valeurs *matepairs* élevées indiquent des sous-groupes bien homogènes et séparés validant la classification tandis qu'un seuil *matepair* plus faible entraîne plus de classes. Si les *matepairs* sont au-dessus d'un certain seuil, les deux sous-groupes sont ajoutés séparément à la file pour être potentiellement redivisés ultérieurement.En revanche, si au moins une des valeurs de *matepair* est au dessous de ce seuil, les sous-groupes sont considérés comme formant un seul groupe et le groupe initial est sauvegardé comme une famille unique.Si la classification

est valide, les deux sous-groupes sont ajoutés dans la file, sinon le groupe initial est sauvegardé comme une classe unique.

2.2.4 Paramétrage

Faut-il donner explicitement les paramètres ? Je peux donner ceux de l'intra-classification et ceux de la super-classification. -PM :0.90
-pour les grands jeux de données, lda à 100 000 -Taille minimale d'une famille : 100 séquences pour la classification et 20 pour la super-classification

2.3 Alignement, consensus et phylogénie

L'alignement des séquence est fait avec muscle [2]. La phylogénie est reconstruite avec Seaview [3] utilisant la méthode du maximum de vraisemblance (PhyML) [4]. Le modèle F84 est utilisé pour la construction de l'arbre. Le support de branche est aLRT (SH-like), sans bootstrap. La fréquences d'équilibre nucélotidique est optimisée. Le ratio de transition et transversion est fixé à 4. Aucun site est considéré comme invariable. Le taux de variation à travers le site est optimisé. Les opérations de recherche d'arbre est NNI et l'arbre de départ est défini avec la méthode de Neighbor-Joining [5] avec une topologie optimisée. Les consensus sont obtenus avec des scripts développés par l'équipe. Les motifs CENP-B (TTCGTTGGAA[AG]CGGGA), PJα (TTCCTTTT[CT]CACC[AG]TAG) et pKβ (CTATAGGGCCAAAGGAA) ont été identifiés avec le logiciel fuzznuc (package EMBOSS) [6] et en autorisant 2 différences au maximum par rapport au consensus.

3 Résultat

3.1 Caractérisaton des familles dans plusieurs espèces

3.1.1 Identification des familles

Les espèce *C. solatus* et *C. pogonias* sont analysées dans un premier temps pour évaluer la classification automatisée en la comparant avec la classification expérimentale du laboratoire. Expérimentalement, 6 familles α-satellites ont été déterminées chez les *Cercopithèques*. Ces deux espèces partagent deux grandes familles monomériques, C1 et C2, de l'ordre de plusieurs milliers de séquences, et deux familles formant un dimère, C3-C4, de l'ordre d'une centaine de

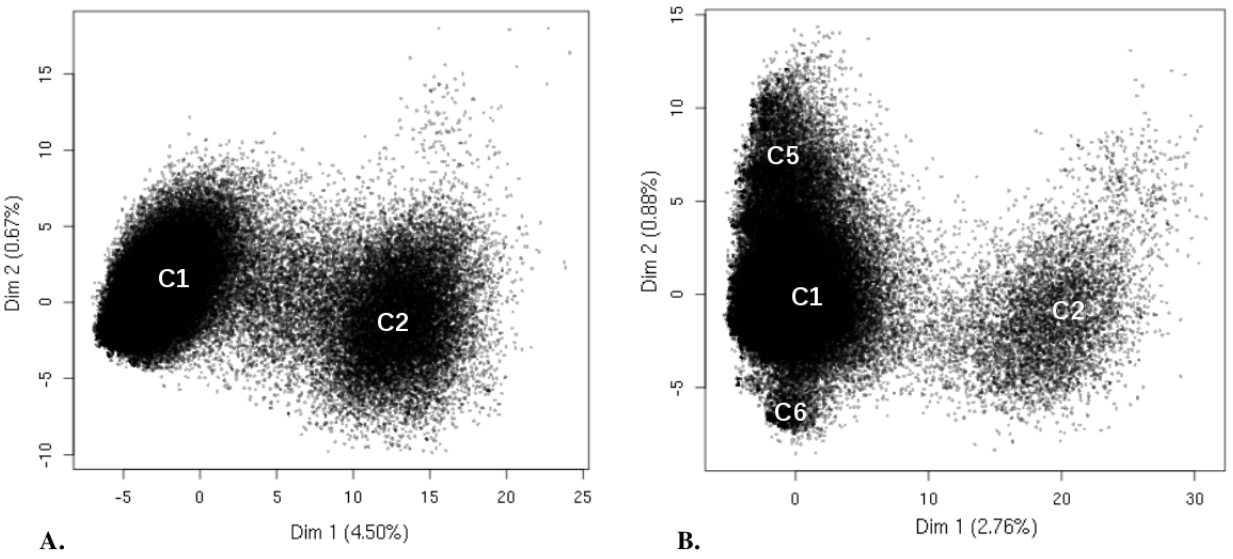

FIGURE 3 – **Caractérisation visuelle des familles α-satellite chez *C. solatus* et *C. pogonias* :** Le nom des familles est indiqué sur les graphiques. Un point représente un monomère. **A.** Les familles présentes chez *C. solatus*. **B.** Les familles présentes chez *C. pogonias*.

séquences chacune. *C. pogonias* possède les familles supplémentaires C5 et C6. Ces familles ont été définies à partir d'une méthode visuelle établie d'après une ACP (Fig. 3).

La classification automatisée donne des familles de taille variable allant de deux à des dizaines de milliers de séquences. Seules les familles ayant plus de 100 séquences, appelées "grandes familles", sont conservées pour l'analyse des α-satellites. Les "petites familles" sont prises en compte en terme de pourcentage de séquences du jeu de données initial qui ne figurent pas dans l'analyse. Les séquences α-satellites chez *solatus* sont réparties en 564 familles, dont 12 grandes familles. Les séquences qui ne sont pas retenues représentent 3,97% du jeu de données. Chez pogonias, le nombre total de familles est de 132, avec 13 grandes familles et 1,29% du jeu de données qui ne figure pas dans les analyses.

Bien que le nombre de grandes familles est relativement proche entre ces deux espèces, les résultats expérimentaux diffèrent significativement. Toutes les familles chez *C. solatus* sont retrouvées : 11 familles forment la famille C2, une famille forme la famille C1 et les familles C3 et C4 sont retrouvées dans des petites familles d'environ 80 séquences chacune. Chez *C. pogonias*, toutes les familles sont retrouvées sauf la famille C6. La famille C1 est répartie en 10 familles, les familles C2, C3 et C5 sont retrouvées entièrement, la famille C4 est également retrouvée sous la

forme d'une petite famille de 86 séquences. Ces résultats contredisent les résultats expérimentaux. La famille C2 est une famille qui a 85% d'identité de séquence, comparé à 95% pour C1. Normalement seule la famille C2 est censée être divisée en plusieurs familles, or ce n'est pas le cas chez *C. pogonias*.

Pour visualiser cette nouvelle répartition des familles, des couleurs sont assignées aux familles issues de la classification automatisée. Ces couleurs sont superposées aux résultats expérimentaux en noir. Chez *C. solatus*, la famille C1 est entièrement retrouvée. La famille C2 est répartie en plusieurs familles. Deux familles intermédiaires, rouge et bleue, sont visibles entre la famille C1, en vert, et C2, en orange. Elles ne semblent pas distinctes. Une famille supplémentaire est retrouvée en turquoise, et se démarque. Pour confirmer cette division de la famille C2, la visualisation de l'ACP des 5-mers est observée en fonction des composantes 1 et 3. Les familles intermédiaires restent emmêlées, contrairement à la famille turquoise qui forme une famille à part entière. Pour certifier ce fait, l'arbre construit atteste que chaque famille est bien retrouvée, notamment les familles intermédiaires qui forment bien deux familles. Chez *C. pogonias*, les familles C2, C4 et C5 sont bien retrouvées. La famille C6 se fond dans la famille C1 (en vert). La famille C1 est divisée en deux familles supplémentaires visibles en rose et violet. La visualisation des composantes 1 et 3 de l'ACP ne permet pas de trancher sur la classification. L'arbre montre que les familles C1 en rose et violet sont très proches, et la famille C4 se démarque.

Le *C. sabaeus* a 338 familles au total, dont 44 grandes familles, et 10,89% du jeu de données qui n'est pas pris en compte. Le *M. fascicularis* a respectivement 709 et 998 familles, dont 42 et 81 grandes familles, et 14,56% et 5,05% du jeu de données qui n'est pas pris en compte. Ces espèces ont beaucoup plus de grandes familles que les *Cercopithèques*, et le deuxième jeu de données possède deux fois plus de grandes familles.

3.1.2 Motifs CENP-B, pJα et pKβ

La protéine CENP-B est présente chez toutes les espèces, mais les *Cercopithèques* ne possèdent pas son site de liaison. Cette assertion est vérifiée par l'absence de pourcentage pour ce motif dans les graphes. *C. sabaeus* et *M. fascicularis* n'ont pas ce motif non plus.

Au contraire, la protéine pJα est très présente chez *C. solatus*, en moyenne a 90%, et *C. pogonias*, en moyenne à 85%. La plus grande famille, C1, ayant plus de 80 000 séquences, se démarque avec un pourcentage à 95%. *C. sabaeus* et *M. fascicularis* ont des pourcentages pour le

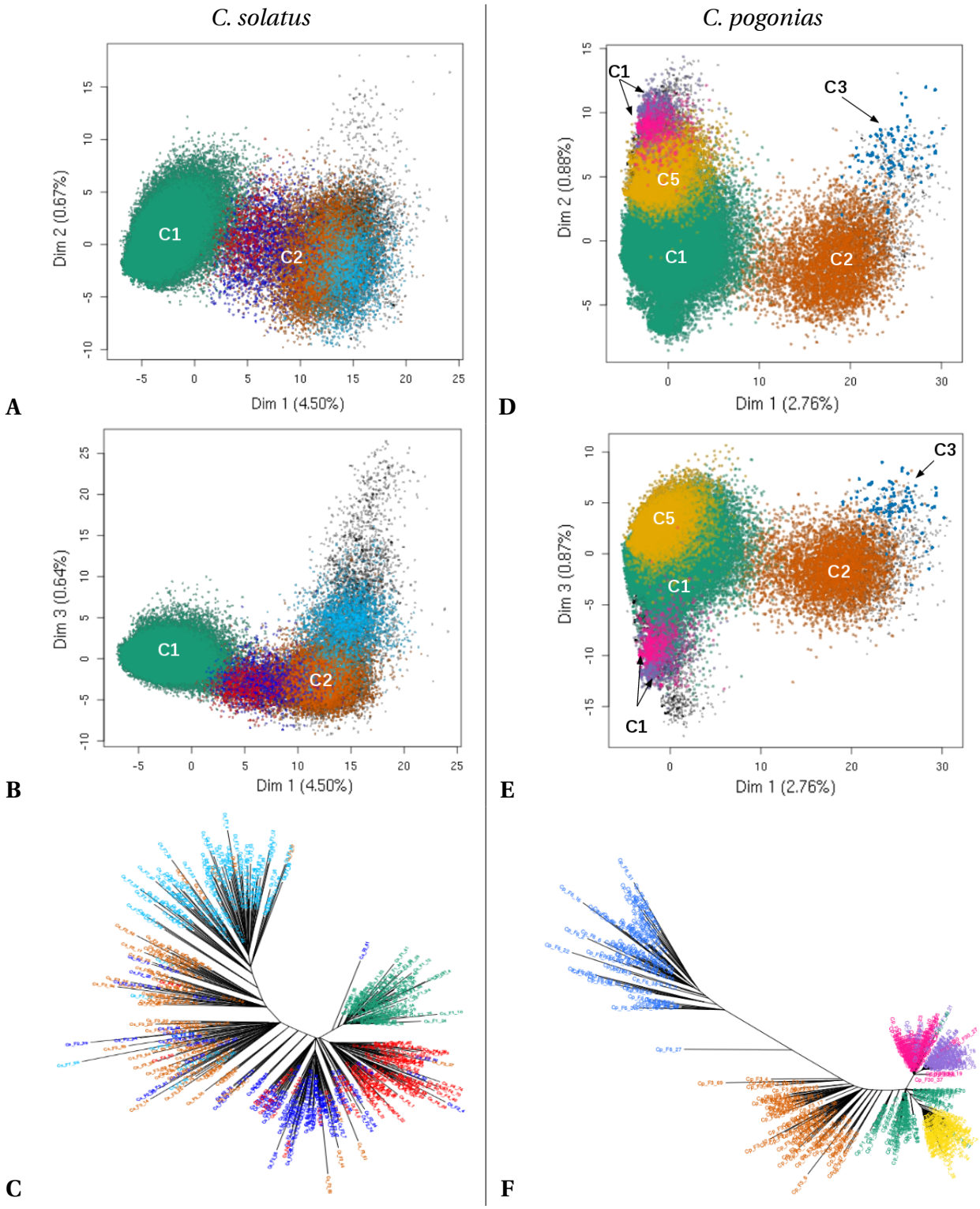

FIGURE 4 – **Les plus grandes familles issues de la classification automatisée :** Les familles expérimentales retrouvées par la classification automatisée sont indiquées sur les représentations de l'ACP des 5-mers. **A.** Composantes 1 et 2 de l'ACP. C1 est en vert, C2 est en orange, rouge, bleu et turquoise. **B.** Composantes 1 et 3 de l'ACP. **D.** Composantes 1 et 2 de l'ACP. De même pour *C. pogonias*. C1 est en vert, violet et rose, C2 est en orange, C4 est en bleu clair, C5 est en jaune. **E.** Composantes 1 et 3 de l'ACP. **C. et F.** Représentation des familles issues de la classification sous forme d'arbres. Les couleurs des familles sont respectivement conservées et 100 séquences alétoires par familles sont sélectionnées pour la construction de l'arbre.

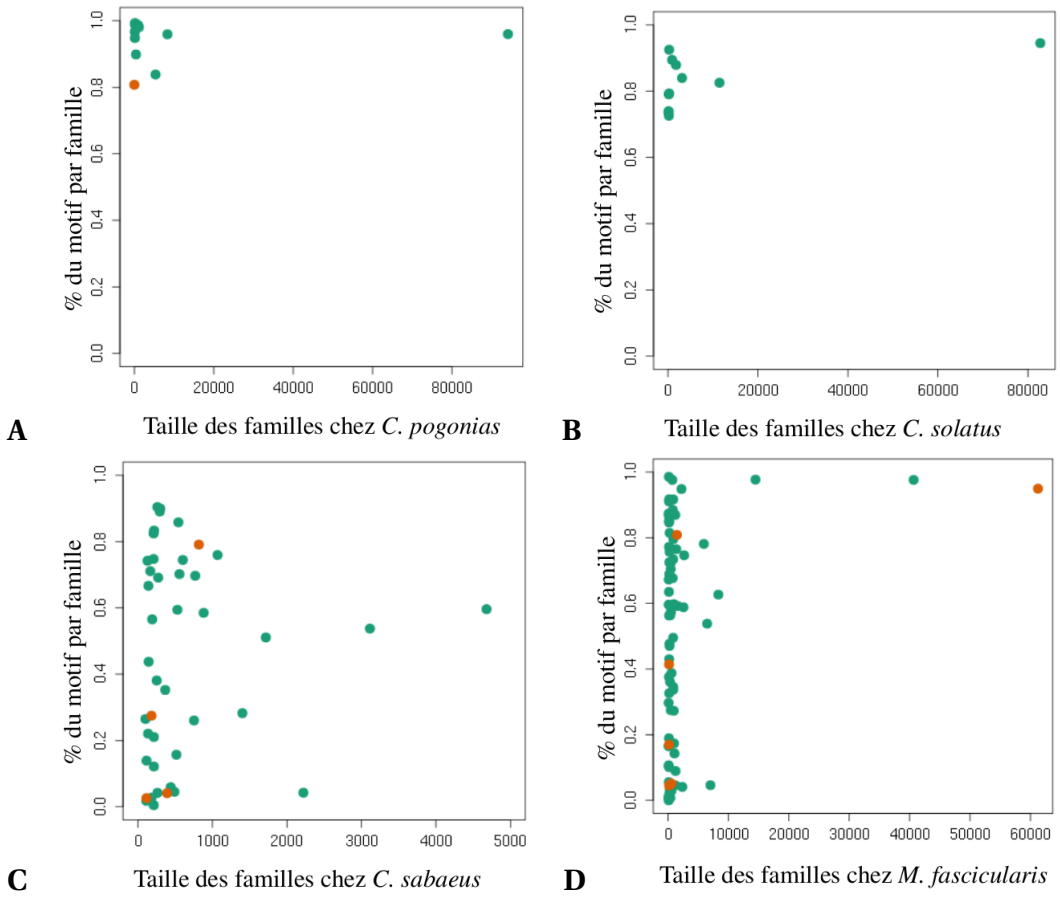

FIGURE 5 – **Présence des motifs CENP-B, pJα ou pKβ par famille :** Le pourcentage de séquences par famille ayant le motif pJα est en vert, pKβ en orange et CENP-B en bleu. Chaque famille est représentée en fonction de sa taille.

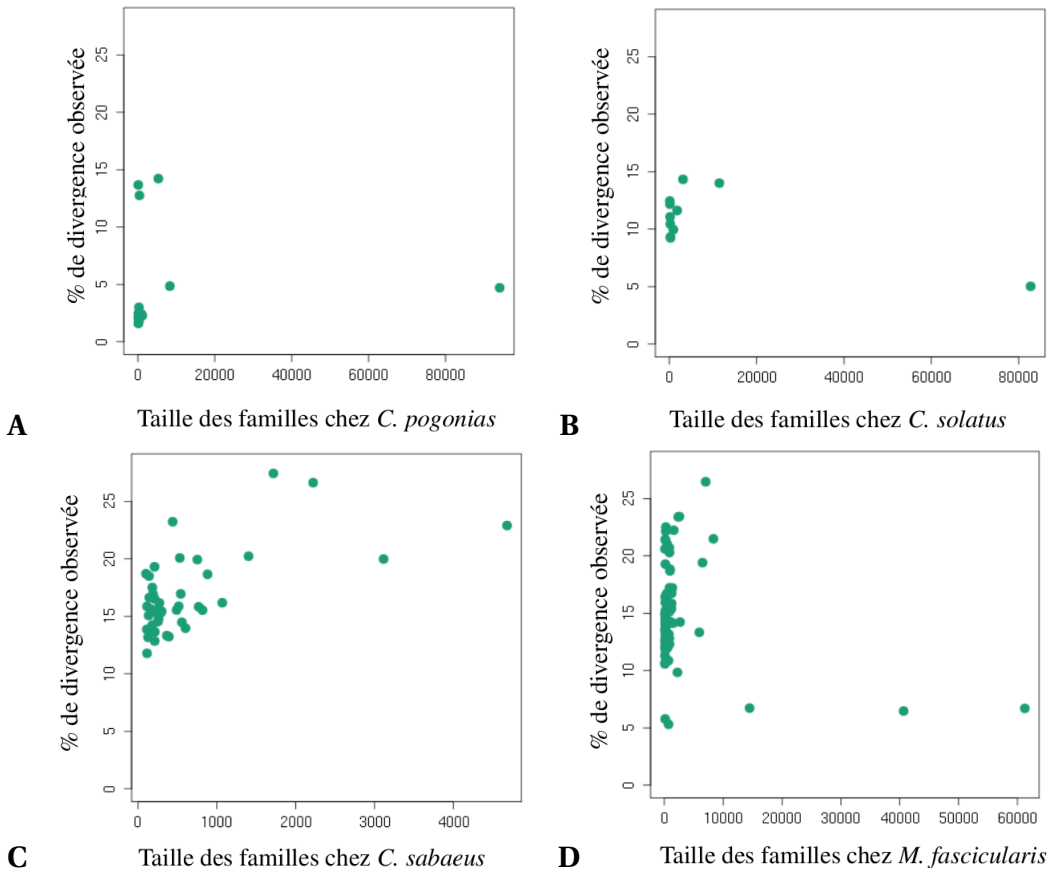

FIGURE 6 – **Pourcentage de divergence observée au sein d'une famille :** Un point correspond à une famille et représente le pourcentage de divergence en fonction de la taille de la famille.

motif pJα qui varie entre 1% et 97%. Parmi ces familles, *C. sabaeus* a une grande famille de 2 200 séquences qui possède très peu ce motif. Le *M. fascicularis* deux grandes familles, de 20 000 et 40 000 séquences ont ce motif à 97%. Le motif pJ$alpha$ est bien préservé chez les primates.

Le motif pKβ est présent lorsque pJα est absent de la famille. Il est absent chez *C. pogonias*. Une seule famille a ce motif à 80% chez *C. solatus*, la famille C3, et chez *C. sabaeus* . *M. fascicularis* a deux grandes familles de plus de 10 000 séquences et plus de 60 000 séquences qui ont respectivement 80% et 95% le motif. Ce motif

Une hypothèse supposerait que les familles ayant un grand pourcentage de pJα sont des familles apparues récemment. Les familles les plus anciennes auraient accumulé plus de mutations à cet endroit, réduisant ainsi le pourcentage de ce motif.

3.1.3 Similarité entre familles

C. solatus et *C. pogonias* ont un pourcentage de divergence qui est relativement faible qui ne dépasse pas les 15%. La famille C1 a un pourcentage de divergence très faible de 5%. Cette

grande famille de plus de 80 000 séquences est donc probablement une famille récente. *C. pogonias* a d'autres familles qui ont un pourcentage en dessous de 5%. *C. sabaeus* a les pourcentages de divergence les plus élevés, allant jusque 28%, quelque soit la taille de la famille. Cette espèce a des familles anciennes d'α-satellites ou alors cette divergence est le résultat d'erreurs de séquençages. Le *M. fascicularis* a des poucentages de divergence qui varient entre 5% et 27%. 3 grandes familles ayant plus de 10 000 séquences ont un pourcentage de divergence qui se rapproche de 5%. Ces familles sont probablement récentes.

Taille moyenne des consensus :
Mais ça revient à quoi de savoir ça ?
-c. pogonias : 184
-c. solatus : 206
-c. sabaeus : 203
-m. fascicularis : 208

Transition ? ? -

3.2 Comparaison inter-espèce et mécanismes d'évolution

3.2.1 Répartition des super-familles

Pour étudier les mécanismes d'évolution des α-satellites, une classification inter-espèce permet de comprendre la différences. Pour chaque espèce, 100 séquences par famille sont tirées aléatoirement. La classification est effectuée sur ce jeu de données composées de toutes les grandes familles α-satellites de chaques espèce. La taille minimale d'une famille est fixée à 20 séquences. Le terme utilisé pour décrire cette classification inter-espèce est "super-classification".

3.2.2 Mécanisme d'évolution des familles ayant pKβ

Parmi les familles α-satellites, toutes celles qui ont le motif pKβ sont sélectionnées pour regarder comment elles s'assemblent durant la super-classification.

4 Discussion

5 Conclusion

Références

[1] Florence Jornod. Développement d'une méthode de classification pour les séquences répétées centromériques de primates. Master's thesis, Université Paris Diderot - Paris 7, 2016-2017.

[2] Robert C. Edgar. Muscle : multiple sequence alignment with high accuracy and high through-put. *Nucleic acids research*, 32(5) :1792–1797, 2004.

[3] Olivier Gascuel Manolo Gouy, Stéphane Guindon. Seaview version 4 : a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution*, 27(2) :221–224, 2009.

[4] Olivier Gascuel Stéphane Guindon. Estimating maximum likelihood phylogenies with phyml. *Bioinformatics for DNA Sequence Analysis*, 537 :113–137, 2009.

[5] Masatoshi Nei Naruya Saitou. The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4) :406–425, 1987.

[6] Alan Bleasby Peter Rice, Ian Longden. Emboss : the european molecular biology open software suite. 2000.

Résumé

Votre résumé commence ici... ...

Abstract

Abstract begins here... ...